Journal of KIIT. Vol. 20, No. 6, pp. 19-32, Jun. 30, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.6.19

BERT 기반의 사전 학습 언어 모형을 이용한 한국어 문서 추출 요약 베이스라인 설계

박재언*[1], 김지호*[2], 이홍철**

Designing Baseline for Korean Document Summarization using BERT-based Pre-trained Encoder

Jae-Eon Park*[1], Ji-Ho Kim*[2], and Hong-Chul Lee**

본 연구는 4단계 두뇌한국21에 의해 지원되었습니다.

요 약

디지털 문서가 기하급수적으로 증가한 현대 사회에서 문서 내 중요한 정보를 효율적으로 획득하는 것은 중요한 요구사항이 되었다. 그러나 방대한 디지털 문서의 양은 개별 문서의 중요 정보를 식별하고 축약하는 데 어려움을 야기하였다. 문서 요약은 자연어 처리의 한 분야로서 원본 문서의 핵심적인 정보를 유지하는 동시에 중요 문장을 추출 또는 생성하는 작업이다. 하지만 벤치마크로 사용하기에 적절한 한국어 문서 데이터의 부재와 베이스라인 없이 문서 요약 연구가 진행되어 발전이 미진한 상황이다. 본 논문에서는 데이터에 대한 검증과 접근성을 충족하고 글의 특성이 다른 두 개의 문서 집합을 선정하였다. BERT 기반의 다국어 및 한국어 사전 학습 언어 모형들을 선정하여 비교 및 실험하였다. 주요 결과로는 한국어 사전 학습 언어 모형이 ROUGE 점수에서 다국어 사전 학습 언어 모형을 능가하였으며, 이에 대한 원인을 추출된 요약 문장의 비율을 통해 분석하였다.

Abstract

In modern society, where digital documents have increased exponentially, it is essential to efficiently obtain important information within documents. However, due to the vast amount of digital documents, it has become difficult for humans to abbreviate important information on individual documents. Document summarization is a Natural Language Processing field that extracts or generates meaningful sentences shorter than the original document while maintaining key information on the original document. However, since there is no appropriate Korean summarization data for benchmark, research has been conducted without a baseline, and development in this field is insufficient. In this paper, two document datasets that satisfy the accessibility and verification of summarization data and different text characteristics were selected. In addition, BERT-based multilingual and Korean pre-trained language models were selected, compared, and tested. For Korean documents, the Korean pre-trained language models outperformed the multilingual pre-trained language models in ROUGE scores. The cause was analyzed through the extraction ratio of selected summary sentences.

Keywords

deep learning, natural language processing, Korean document summarization, extractive summarization, automatic evaluation metric

* 고려대학교 산업경영공학부
- ORCID[1]: https://orcid.org/0000-0002-0410-6457
- ORCID[2]: https://orcid.org/0000-0003-3733-8702
** 고려대학교 산업경영공학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-4407-0348

· Received: Mar. 10, 2022, Revised: Apr. 07, 2022, Accepted: Apr. 10, 2022

· Corresponding Author: Hong-Chul Lee
Dept. of Industrial and Management Engineering, Korea University, Korea
Tel.: +82-2-3290-3767, Email: hclee@korea.ac.kr

Ⅰ. 서 론

인터넷의 발달로 사람들은 디지털화된 문서에 대해 접근이 용이해졌으며 손쉽게 원하는 정보를 얻을 수 있게 되었다. 이처럼 수많은 정보 가운데서 어떻게 하면 정보를 효과적이고 효율적으로 획득할 수 있을지가 매우 중요한 시대가 되었다. 이를 위해 많은 기술이 제안되었으며 문서 요약이 그중 하나다[1]. 문서 요약은 개별 단어와 문장의 의미를 넘어서 문서 전체의 내용을 이해하고 중요한 문장을 파악하는 자연어 처리 분야의 핵심 과업이다. 요약하려는 원본 문서가 가지는 의미는 유지하는 동시에 문서의 복잡도는 줄이고 원본 문서보다 길이가 짧은 문서를 추출 또는 생성하는데 주목적을 가진다.

문서 요약은 추출 요약(Extractive summarization)과 생성 요약(Abstractive summarization)의 두 가지 방법이 존재한다[2]. 추출 요약은 원본 문서의 문장들 가운데 원본 문서의 중요한 정보를 내포하고 있는 문장 몇 가지를 선택하여 요약문으로 사용하는 방법론이며, 생성 요약은 원본 문서의 문맥을 이해하고 이를 바탕으로 원본 문서에 없는 단어 또는 구로 이루어진 요약문을 생성하는 방법론이다[3]. 하지만 생성 요약은 제약이 없다는 특징으로 인해 결과가 신뢰할 수 없는 경우가 종종 발생한다[4]. 일부 논문들은 좋은 생성 요약문을 생성하고자, 생성 요약문을 만들기 전에 추출 요약으로 얻은 문장 정보를 활용하는 방법을 사용하기도 한다[5][6]. 이처럼 추출 요약은 복잡하지 않고 문법적으로나 의미적으로 정확한 요약을 생성하는 경우가 대부분이기 때문에 중요한 연구 분야이다[7].

기존의 문서 요약 연구들은 정해진 데이터 없이 연구자가 직접 웹 크롤링을 활용하여 뉴스 기사와 같은 문서 데이터셋을 확보하였다[6][8]. 문서 추출 요약에 대한 베이스라인(Baseline)을 설정하기 위해서는 공공이 접근할 수 있고 검증이 된 데이터 세트를 선정하는 것이 중요하다. 하지만 이러한 데이터 세트가 대규모로 구축된 영어와 달리 한국어로 된 요약 데이터 세트는 현재까지 미진한 상황이다[9]. 하지만 최근 과학기술정보통신부 산하의 한국지능정보사회진흥원을 필두로 대규모 인공지능 학습 데이터 세트 구축 사업이 진행되었고, 그 결과 컴퓨터 비전, 음성/자연어, 헬스케어 등의 분야에 접근성과 검증성을 완비한 대규모 학습 데이터 세트가 생성되었다. 본 연구에서는 한국진흥정보사회진흥원의 AI hub 문서 요약 데이터 세트를 활용하였으며, 그 가운데 글의 쓰기 방식, 목적, 요약 방법이 서로 다른 두 개의 단일 문서 데이터를 선정하여 결과를 얻었다.

최근 사전 학습된 언어 모형이 감성 분석(Sentiment analysis), 질의응답(Question & answering), 개체명 인식(Named entity recognition) 등 NLP(Natural Language Processing, 자연어처리)의 핵심적인 연구 분야에서 우수한 성능을 보여주고 있다. 특히 사전 학습된 트랜스포머 기반 인코더(BERT, Bidirectional Encoder Representation from Transformers)의 우수함이 입증되며[10] 주요 연구 분야의 기준을 제공하고 있다. 본 논문에서는 한국어 문서 추출 요약을 위해 여러 가지 다국어 및 한국어 사전 학습 모형을 구축하여 성능을 비교하였다. 비교를 위해 선택한 다국어 사전 학습 언어 모형인 MBERT(Multilingual BERT)[11]와 XLM-R(Cross-lingual Language Model-RoBERTa)[12]은 각각 104개 국어와 100개 국어로 학습이 이루어졌으며, 수많은 언어로 학습이 이루어졌음에도 불구하고 서로 다른 언어 간의 일반화가 가능함을 보여주었다[11][12]. 한국어로 학습된 사전 학습 언어 모형은 SKT의 KoBERT, ETRI의 KorBERT, KLUE의 BERT 등 다수가 존재한다. 본 연구에서는 최근 한국어 NLP의 주요 연구 분야에 대한 벤치마크(Benchmark) 데이터 세트를 구축한 KLUE의 BERT와 RoBERTa를 한국어 사전 학습 언어 모형으로 선정하여 베이스라인을 구축했다.

Ⅱ. 관련 연구

2.1 사전 학습 언어 모형

사전 학습 언어 모형[13]–[15]은 NLP 과업의 광범위한 분야에서 우수한 성능을 보이며 핵심적인 기술로 자리 잡고 있다. 사전 학습 모형이 뛰어난

성능을 보일 수 있었던 이유는 거대한 말뭉치를 바탕으로 문맥적 표현을 학습하였기 때문이다[10]. 이미 학습된 모델을 이용하면 NLP downstream task를 효율적으로 사용할 수 있다[15].

[15]는 RNN(Recurrent Neural Network, 순환신경망)과 같은 이전의 언어 모형은 단방향으로만 토큰을 고려하는 인코딩 방법에 한계가 있음을 주장하였고, 이를 보완하고자 양방향의 BERT를 제안하였다. 기존 RNN계열의 모델은 인코더에서 직전 단어에 집중하여 컨택스트 벡터를 만들기 때문에 장기의존성(Long-term dependency) 반영에 어려움이 있었기 때문이다. 또한, 사전 학습기반의 임베딩 방법론인 ELMo[16]의 경우도 상단 Bi-LSTM 레이어는 양방향이긴 하지만 중간 레이어가 단방향으로 설계되었다. 이와 달리 BERT는 개별 단어의 연산을 위해 모든 단어의 정보를 사용하며 중요 정보에 집중한다는 장점을 가진다.

BERT는 모형에 사용된 구조의 복잡도에 따라 Large와 Base로 나뉜다. 본 연구에서 활용한 Base 모형의 Encoder는 총 12개의 블록, 각 블록의 은닉층의 수는 768개, Self-attention의 수는 12개로 이루어져 있다. BERT는 기존의 Position encoding을 사용하지 않고 Position embedding을 활용한다. RNN과 같은 순환 신경망은 토큰의 순서를 판단할 수 있지만, Transformer는 순환 신경망의 구조를 사용하지 않기 때문에 Position embedding을 통해 순서에 대한 정보를 모형에 알려준다. 이와 더불어 개별 단어에 대한 정보를 가지는 Token embedding과 문장 구분에 대한 정보를 가지는 Segment embedding을 추가하여 Transformer의 입력으로 투입한다. 결합된 임베딩 정보들은 블록 내부로 들어가 Multi-Head Attention을 수행 후 Position-wise FFNN(Feed-Forward Neural Network)을 통과하여 토큰 단위로 출력을 산출한다. 그리고 산출된 토큰 벡터는 입력으로 들어간 문장의 풍부한 문맥적 특징들을 담고 있다.

RoBERTa(A Robustly optimized BERT pretraining approach)[17]는 BERT의 두 메커니즘인 MLM (Masked Language Model, 마스크 언어 모델)과 NSP(Next Sentence Prediction, 다음 문장 예측)에 의문을 제기한다. MLM은 입력으로 들어가는 토큰 가운데 무작위로 15%를 선택하고 그중 80%는 마스킹, 10%는 랜덤 토큰으로 대체, 10%는 그대로 유지하여 이후에 마스킹한 토큰을 예측하는 메커니즘이며, NSP는 두 문장을 입력으로 투입할 때 확률적으로 50%는 기존 문장과 이어지는 후속 문장을, 나머지 50%는 문서 내의 다른 문장을 연결하여 두 문장 간의 관계성을 학습하는 메커니즘이다. 하지만 RoBERTa의 저자는 BERT의 MLM은 전처리 단계에서 고정시키는 정적 마스킹 과정을 거치며, NSP는 두 문장만을 고려하기 때문에 불합리하다고 주장한다. 따라서 정적 마스킹 방법이 아닌 학습할 때마다 마스킹의 위치가 변하는 동적 마스킹 방법으로 대체한다. 또한, NSP는 제거하여 입력 토큰의 길이가 512가 넘지 않는 선에서 최대한 많은 문장을 이어 붙여 모형의 입력으로 투입하는 방법을 택하였다. 이렇게 설계된 RoBERTa는 GLUE benchmark 데이터세트에 대해서 state-of-the-art를 달성하였다.

2.2 추출 요약

추출 요약은 문서 내에서 가장 중요한 문장을 식별하여 요약문을 만드는 방법이다. 문장을 식별하는 과정에서 신경망 모형은 개별 문장에 대해 표현을 학습하고 학습된 문장 표현을 바탕으로 어떤 문장을 요약문으로 사용할지를 예측하는 지도학습 과업으로 간주한다[10]. 추출 요약을 위한 가장 초기 모형은 [7]로 매 타임스탭마다 문장 내의 단어를 입력으로 받아 양방향 순환 신경망으로부터 단어에 대한 벡터값을 획득 후 평균을 내어 문장을 표현하였다. 모든 문장에 대해 벡터값을 얻은 후 다시 순환 신경망을 통과하여 요약문에 사용할 문장을 판단한다. [18]는 ROUGE 지표를 최적화하는 새로운 훈련 알고리즘을 통해 추출 요약을 문장의 순위를 매기는 강화학습으로 치환하였다. [19]는 추출 요약의 문장별로 점수를 내고 점수를 바탕으로 요약 문장을 선택하는 두 단계 과정을 하나로 통합한다. 매 타임스탭마다 해당 문장이 요약문에 포함될지 여부를 ROUGE 점수에 기반한 목적함수를 통해 판단한다.

III. 베이스라인 설계 방안

이 논문에서는 앞서 기술한 사전 언어 학습 모형을 활용하여 베이스라인을 설계한다.

그림 1은 베이스라인 설계의 전체적인 개요를 나타내며 데이터 수집, 데이터 전처리, 모델 선정 그리고 선정된 모델에 대한 성능 분석으로 구성된다. 단계별 자세한 설명은 다음과 같다.

3.1 데이터 수집

AI hub 문서 요약 데이터 세트는 서로 다른 양식의 문서 세 개로 구성되어 있다. 그 가운데 글의 작성 방식과 목적이 다른 두 개를 선정하였다. 신문 기사는 10개 언론사로부터 30만 건의 데이터로 구축되었으며, 이 중 종합면 30%, 정치 20%, 경제 20%, 사회 20%, 문화 및 스포츠 기타 10%의 비율로 구성되어 있다. 사설/잡지는 신문의 사설면을 통해 확보한 기고문과 시사/경제, 공학/기술 등 7개의 카테고리의 전문적인 내용을 포함한 7만 건의 데이터로 구축되었다[9].

3.2 데이터 전처리

신문 기사의 전처리는 [20]의 방법을 적용하였다. 우선 한글이 아닌 특수한 문자를 제거하였으며, 신문 기사에서 흔히 볼 수 있는 기자의 이름, 이메일, 언론, 사진, 출처 등을 제거하였다. 또한, 언어 모형이 편향되지 않도록 하고자 사전 학습된 모델1 (https://huggingface.co/monologg/koelectra-base-v3-gender-bias), 모델2(https://huggingface.co/monologg/koelectra-base-v3-hate-speech)을 이용하여 차별 및 혐오적 문장을 포함하는 일부 기사를 제외하였다.

사설/잡지의 학습 데이터의 문서 소분류에는 잡지 문서가 존재하지만, 검증 데이터에는 존재하지 않는 문제가 존재하였다. 언어 모형의 편향성을 줄이기 위해 학습 데이터와 검증 데이터를 합친 후 소분류를 기준으로 층화 추출을 통해 잡지 문서가 학습 데이터와 검증 데이터에 8:2의 비율로 속하도록 재가공하였다. 재가공을 거친 사설/잡지의 학습 데이터와 검증 데이터는 신문 기사와 동일한 전처리 과정을 거쳤다. 전처리가 완료된 신문 기사와 사설/잡지 데이터의 통계량은 표 1과 같으며 전처리 결과에 대한 예시는 표 2와 같다.

표 1. 실험 데이터 세트 (Docs: Documents, Sents: Sentences)
Table 1. Experimental data sets (Docs: Documents, Sents: Sentences)

Dataset	# of docs (train/valid/test)	Avg. Doc length		Avg. summary length	
		words	sents	words	sents
Newspaper	264,088/14,949 /15,061	232.60	13.80	57.82	3.0
Editorial/ magazine	53,616/6,714 /7,008	270.64	20.83	43.59	3.0

3.3 모델 선정

Baseline을 설정하기 위해서는 동일한 구조를 가진 모형에 대한 선정이 필요하다. 따라서 본 연구에서는 BERT 기반의 Base model만을 선정하여 실험을 진행하였다. BERT의 Base model의 Encoder는 총 12개의 블록, 각 블록의 은닉층의 수는 768개, Self-attention의 수는 12개로 이루어져 있다

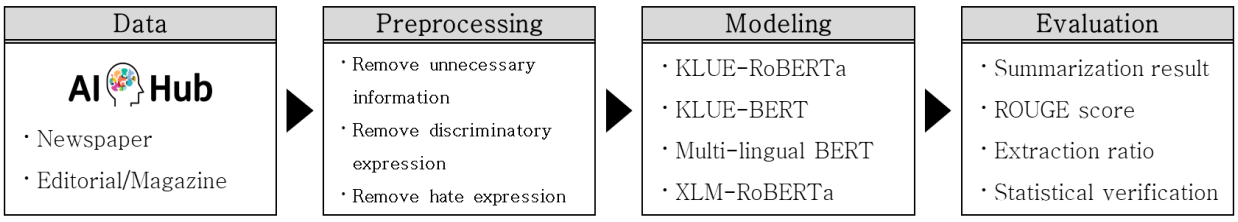

그림 1. 설계 방안의 전반적인 프로세스
Fig. 1. Overall process of designing method

표 2. 전처리 결과 예시
Table 2. Example of data preprocessing result

Before preprocessing	총 12개소 대상, 위험요소 사전 차단. 김영신 기자 yskim0966@naver.com. 광양시는 광양소방서와 함께 시민들이 많이 이용하고 있는 다중이용시설 중 화재에 취약한 건축물을 대상으로 소방 특별점검을 실시했다. 이번 소방 특별점검은 화재가 발생해 인명피해가 발생한 제천 노블휘트니스 스파와 유사한 건물 3개소와 인명피해가 우려되는 다중이용시설 9개소 등 총 12개소를 대상으로 진행됐다. 소방공무원과 관련 전문가, 공무원으로 구성된 점검반은 지난 3일부터 화재 시 피난할 수 있는 비상구 폐쇄 및 비상구나 피난통로에 장애물 설치 여부, 소방시설 정상작동 여부와 관리 상태 등 화재위험과 인명피해 우려되는 요소들을 집중적으로 점검했다. 조사결과 소방시설 불량과 건축물 임의 증축, 비상구 다른 용도 활용 등으로 적발된 곳에는 현지 시정명령 등의 조치를 취했으며, 소방서에서도 소방시설 불량사항에 대한 조치명령 발부와 취약대상 소방시설을 대상으로 소방훈련을 실시할 계획이다. 조춘규 안전총괄과장은“다중복합시설은 내부 구조가 복잡해·화재가 발생하면 연기로 인해 비상구를 찾기 매우 어렵다”며“건물 관리자는 피난통로의 장애물을 제거하고 소방시설 정상 작동을 확인하고, 이용자는 비상구를 미리 확인하는 등 모두가 관심을 가져야 인명피해를 막을 수 있다”고 당부했다. 한편, 소방당국은 제천 화재사고와 유사 건물 12개소 외에도 일반 숙박시설 등을 지속적으로 점검해 화재 발생에 강력히 대응해 나갈 방침이다.
After preprocessing	총 12개소 대상 위험요소 사전 차단. 광양시는 광양소방서와 함께 시민들이 많이 이용하고 있는 다중이용시설 중 화재에 취약한 건축물을 대상으로 소방 특별점검을 실시했다. 이번 소방 특별점검은 화재가 발생해 인명피해가 발생한 제천 노블휘트니스 스파와 유사한 건물 3개소와 인명피해가 우려되는 다중이용시설 9개소 등 총 12개소를 대상으로 진행됐다. 소방공무원과 관련 전문가 공무원으로 구성된 점검반은 지난 3일부터 화재 시 피난할 수 있는 비상구 폐쇄 및 비상구나 피난통로에 장애물 설치 여부 소방시설 정상작동 여부와 관리 상태 등 화재위험과 인명피해 우려되는 요소들을 집중적으로 점검했다. 조사결과 소방시설 불량과 건축물 임의 증축 비상구 다른 용도 활용 등으로 적발된 곳에는 현지 시정명령 등의 조치를 취했으며 소방서에서도 소방시설 불량사항에 대한 조치명령 발부와 취약대상 소방시설을 대상으로 소방훈련을 실시할 계획이다. 조춘규 안전총괄과장은 다중복합시설은 내부 구조가 복잡해 화재가 발생하면 연기로 인해 비상구를 찾기 매우 어렵다며 건물 관리자는 피난통로의 장애물을 제거하고 소방시설 정상 작동을 확인하고 이용자는 비상구를 미리 확인하는 등 모두가 관심을 가져야 인명피해를 막을 수 있다 고 당부했다. 한편 소방당국은 제천 화재사고와 유사 건물 12개소 외에도 일반 숙박시설 등을 지속적으로 점검해 화재 발생에 강력히 대응해 나갈 방침이다.

표 3. 사전 학습 언어 모형 비교 (lgs: languages)
Table 3. Comparison of pre-trained language models (lgs: languages)

Model	# of lgs	Embedding size	Hidden size	# of layers	# of heads	# of parameters	Vocab size
(a) KLUE-BERT	1	768	768	12	12	110M	32K
(b) MBERT-cased	104	768	768	12	12	172M	110K
(c) KLUE-RoBERTa	1	768	768	12	12	125M	32K
(d) XLM-R	100	768	768	12	12	270M	250K

(a) https://huggingface.co/klue/bert-base, (b) https://huggingface.co/bert-base-multilingual-cased,
(c) https://huggingface.co/klue/roberta-base, (d) https://huggingface.co/xlm-roberta-base

또한, 한국어 문서 요약에 대한 적합한 모델을 고르기 위해 한국어 사전 학습 언어 모형과 다국어 사전 학습 언어 모형 각각 2개를 선정하여 실험을 진행하였다. 각 모델들의 구체적인 구조에 대한 비교는 표 3과 같다.

3.4 성능 평가

학습된 언어 모형이 추출한 문장들로 구성된 요약문에 대한 성능 평가는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)를 사용하여 평가하였다. ROUGE의 구체적인 수식은 다음과 같다.

$$Rouge-N= \frac{\sum_{S \in Reference\ summaries} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in Reference\ summaries} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

n은 N-gram($gram_n$)의 길이를 의미하며, $Count_{match}(gram_n)$은 정답 요약문과 언어 모형이 생성한 문장 간의 N-gram으로 겹치는 단어의 최대 개수를 의미한다[21]. ROUGE-1과 ROUGE-2는 각각 Unigram(1-gram)과 Bigram(2-gram)으로 계산된 지표이며 구체적인 정보가 포함되었는지를 평가한다. 반면 ROUGE-L은 LCS(Longest Common Sequence) 알고리즘을 적용하여 최장 길이로 겹치는 문자열을 계산한다. LCS 알고리즘은 문자열이 연속적이지 않더라도 중복이 되는 정보를 계산하므로 정보가 풍부하게 포함되는지를 판단할 수 있다.

ROUGE 점수는 전반적인 언어 모형의 성능을 보는 평가 방법이라면, 추출 비율은 실제 언어 모형이 문서 내의 어느 위치의 문장을 집중적으로 선택하여 요약문으로 채택했는지를 판단할 수 있다. 이를 통해 언어 모형의 학습이 어떻게 이루어졌는지를 가늠할 수 있다.

3.5 추출 요약

추출 요약의 프로세스는 다음과 같다. M개의 문장($s_1,s_2,...,s_M$)으로 이루어졌으며 문장 s_i는 N개의 단어($w_1,w_2,...,w_N$)으로 구성된 문서 D가 주어졌다고 가정하자. 추출 요약은 문서 D로부터 m개의 문장($m \leq \mathrm{M}$)을 선택함으로써 요약문 S를 만들어 내는 것을 목표로 한다. 문서 내의 각 문장($s_i \in \mathrm{D}$)에 대하여 정답 값 $y_i \in 0,1$이 존재하며, 새로운 문서 D'의 문장 M'에 대하여 $\hat{y} \in 0,1$을 예측한다. 최종적으로 1로 예측된 문장 s_i는 요약 문장에 포함된다.

3.6 추출 요약을 위한 사전 학습 BERT

BERT[15]는 NLP의 광범위한 분야에서 사용되지만, 핵심적인 두 메커니즘인 MLM과 NSP 때문에 직접적으로는 문서 요약 과업에 사용할 수 없다. MLM으로 인해 BERT의 결과가 토큰으로 나오기 때문에, 첫 문장 이후 어디서부터가 후속 문장인지에 대한 구분이 어려워 문서 요약 과업에 적합하지 않다. 이를 해결하고자 [10]는 각 문장의 시작에 classification 토큰([CLS])를 삽입하였다. 문장의 시작을 알리는 [CLS] 토큰의 존재로 MLM이 가지는 문제를 해결하여, 문장 간의 구분이 가능해졌다. 또한, 여러 문장을 입력으로 받아 문장 간의 관계성을 파악해야 하는 문서 요약에서 두 문장만을 투입받는 BERT의 구조(NSP)가 불합리하다.

이를 위해 [10]은 기존의 Segment embedding이 아닌 Interval segment embedding을 도입하여 입력으로 들어온 두 개 이상의 문장에 대하여 구분을 지을 수 있도록 하였다. Interval segment embedding의 문장 구분 방법은 가령 문서의 문장 [$sent_1$,$sent_2$, $sent_3$,$sent_4$,$sent_5$]이 있을 때, 해당 문장의 순번이 홀수([$sent_1$,$sent_3$,$sent_5$])인지 또는 짝수([$sent_2$, $sent_4$])인지에 따라 E_A 또는 E_B의 임베딩으로 구분하여 할당하므로 문장 간의 관계성을 학습할 수 있다. 기존의 BERT와 마찬가지로 최종 출력은 토큰 단위로 산출된다.

그림 2는 BERT를 활용한 추출 요약 네트워크이다. 추출 요약에 사용되는 Transformer의 내부 Encoder는 기존 BERT에 사용되는 Transformer의 Encoder와 동일한 구조를 가진다. 하지만 추출 요약을 위한 BERT는 기존의 BERT와 가장 상단에서 차이를 보인다. 기존 BERT와는 달리 추출 요약을 위한 BERT는 각 문장 앞에 새롭게 CLS 토큰을 추가하였기 때문에, 가장 상단의 층을 통과하여 출력된 토큰 t_i는 i번째 [CLS] 토큰 벡터를 이용하여 i번째 문장 전체를 표현하게 된다. 최종적으로 i번째 문장을 대표하는 토큰 t_i를 시그모이드 함수(σ)에 투입하여 요약 문장으로 추출되는 여부에 따라 0과 1로 분류한다.

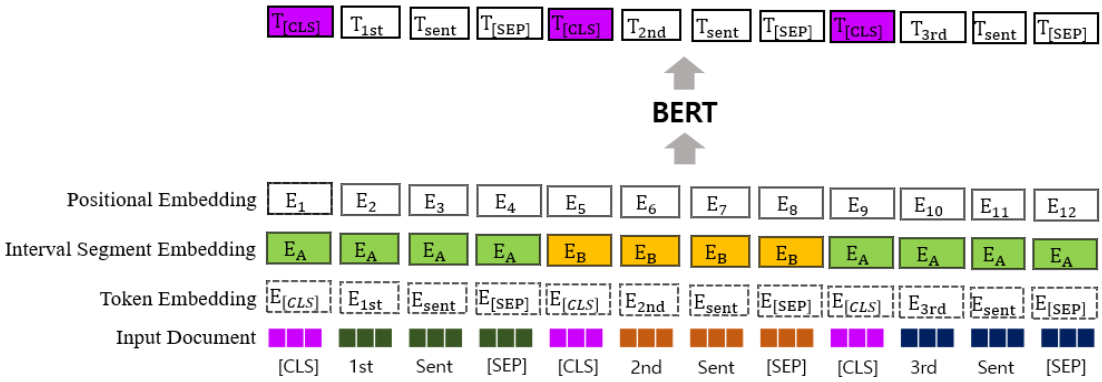

그림 2. [10]의 BERT를 활용한 요약 모형 구조
Fig. 2. BERT summarization architecture using [10]

$$\hat{y}=\sigma\left(Wh_i^L+b\right) \quad (2)$$

여기서 h_i^L은 Encoder의 마지막 L번째 층을 거친 i번째 문장의 벡터값을 나타낸다. 가령, 문서 내의 문장이 $[sent_1,sent_2,sent_3,sent_4,sent_5]$과 같을 때, 추출 여부가 $[1,1,0,1,0]$과 같다면 요약 문장은 $[sent_1,sent_2,sent_4]$로 구성된다.

Ⅳ. 실 험

4.1 실험 환경 및 세팅

모든 실험은 GPU 3개 (NVIDIA Tesla V100)를 이용하여 최대 50,000회를 학습하였다. 50,000회를 학습하는 가운데 1,000회마다 검증 데이터에 대한 평가와 체크 포인트가 저장되었다. 검증 데이터에 대한 손실 값이 가장 작을 때의 체크 포인트 정보를 불러와 테스트 데이터의 결과를 산출하였다. 학습이 끝난 언어 모형에 대한 성능 평가는 ROUGE-1/2/L 점수를 사용하였다[21]. 그중에서도 문서 요약 분야는 ROUGE-2 점수를 기준으로 언어모형의 성능을 평가한다.

추출 요약 과업에서는 언어 모형의 상한 성능을 확인하기 위해 Oracle summary를 계산한다[22]. 각 문서에 대한 Oracle summary를 얻기 위해 ROUGE-2에 대해서 탐욕 알고리즘(Greedy algorithm)을 적용하였으며, Oracle summary를 얻는 과정은 다음과 같다. 문서 내의 문장을 순회하며 정답 요약문(Reference summary)과 비교했을 때, ROUGE-2 점수가 가장 높은 문장을 임시 정답 문장으로 선택한다. 임시 정답 문장이 채택된 후 다시 문서 내의 문장을 순회하며 채택된 문장과 함께 ROUGE-2 점수를 최대화하는 문장을 선택한다. 이를 문서 마지막 문장까지 반복하며 ROUGE-2 점수가 더 상승하지 않으면 종료한다[7]. 추출 요약에서 Oracle summary는 정답 요약문에 대해 생성되므로 언어 모형은 이 점수를 넘어설 수 없다.

손실 함수는 교차 엔트로피로 설정하였고 학습률은 1e-5로 고정 후 AdamW[23]를 최적화 함수로 사용하였다. 테스트 데이터에 대해 추출 요약의 진행 시, 학습이 끝난 언어 모형을 이용하여 문서의 각 문장별로 확률값을 계산한다. 기존의 추출 요약의 설계는 3.1절에서 언급했듯이 시그모이드 함수를 통해 0 또는 1로 요약 문장을 선택한다. 하지만 문서 내의 문장이 많아지면 벡터가 희소(Sparse)해지는 문제가 발생하며 이는 모형 학습에 치명적인 손해를 야기한다. 따라서 시그모이드 함수가 아닌 소프트맥스 함수를 활용하여 문서 내 문장별로 확률값을 계산하였다. 확률값을 산출 후 내림차순 정렬하여 상위 3개의 문장(Top-K strategy)을 선택하였고 이를 바탕으로 요약문을 생성하였다.

4.2 실험 결과

표 4와 표 5는 각각 신문 기사와 사설/잡지 데이터에 대한 실험 결과를 요약한 것이다. 도표의 가장

상단의 블록은 추출 요약에서 언어 모형의 성능을 비교하기 위한 베이스라인으로 사용된다. ORACLE은 4.1절에서 설명했듯이 모형의 상한 성능을 확인하기 위해 사용되는 지표이며, LEAD-3는 문서의 첫 세 문장을 해당 문서의 요약문으로 간주하여 성능을 비교하는 지표이다[3].

두 번째 블록은 한국어와 다국어의 사전 학습 언어 모형의 결과 비교이다. 신문 기사와 사설/잡지 데이터에 대해서 한국어 사전 학습 언어 모형이 다국어 사전 학습 언어 모형보다 우수한 성능을 보였다. RoBERTa 기준으로 신문 기사 테스트 데이터의 결과를 보았을 때, 한국어 사전 학습 언어 모형이 다국어 사전 학습 언어 모형보다 대략 0.7% 가량 우수하였다. 가장 우수한 모형인 KLUE-RoBERTa와 가장 약한 모형인 MBERT-cased의 차이는 약 9.5% 차이가 났다. 사설/잡지 테스트 데이터의 결과 또한 신문 기사의 결과와 유사하였다. 한국어 RoBERTa 사전 학습 모형의 결과가 다국어 RoBERTa 사전 학습 모형의 결과보다 3.4%가량 뛰어났다. 또한, KLUE-RoBERTa가 MBERT-cased보다 약 15.4% 우수한 결과를 보였다.

표 4. 신문 기사 데이터에 대한 ROUGE-2 내림 차순 정렬
Table 4. Descending order of ROUGE-2 for newspaper data

Model	R1	R2	RL
ORACLE	75.90	72.55	75.80
LEAD-3	63.74	60.14	63.64
Extractive			
KLUE-RoBERTa	67.98	64.62	67.89
KLUE-BERT	67.53	64.14	67.45
XLM-R	67.30	63.89	67.22
MBERT-cased	58.50	54.48	58.38

표 5. 사설/잡지 데이터에 대한 ROUGE-2 내림차순 정렬
Table 5. Descending order of ROUGE-2 for editorial/ magazine data

Model	R1	R2	RL
ORACLE	60.81	55.58	60.75
LEAD-3	33.12	29.07	33.04
Extractive			
KLUE-RoBERTa	47.02	42.69	46.93
KLUE-BERT	46.07	41.73	45.98
XLM-R	43.60	39.21	43.50
MBERT-cased	31.60	27.45	31.51

표 3의 모델 간 비교를 보았을 때, KLUE-BERT의 단어 크기가 XLM-R에 비해 대략 2.5배 적음에도 불구하고 ROUGE-1/2/L가 두 문서 데이터에서 앞서는 것을 확인할 수 있다. 이를 통해 한국어 데이터에 대해서는 한국어 사전 학습 언어 모형의 단어 사전 크기가 작더라도 다국어 사전 학습 언어 모형보다 좋다는 결론에 도달하였다. 표 6과 표 7은 각각 신문 기사와 사설/잡지 데이터에 대해서 각 모델이 추출한 요약 문장에 대한 예시이다.

4.3 실험 결과 분석

표 4와 표 5의 다국어 사전 학습 언어 모형인 MBERT의 결과가 LEAD-3보다 낮은 이유는 BERT 네트워크의 한계와 관련이 있다. [17]는 BERT가 상당히 Undertrained되었다고 주장하며 BERT를 개선한 RoBERTa 네트워크를 제안한다. XLM-R는 두 문서 모두에서 LEAD-3보다 우수하다는 점을 바탕으로 RoBERTa 네트워크가 BERT네트워크보다 학습에 이점이 있다는 것을 알 수 있다. 마찬가지로 한국어 사전 학습 언어 모형에 대해서도 동일한 결과를 보였다.

또한, 상당한 단어 사전 크기를 구축하고 더 많은 파라미터를 가진 다국어 언어 모형인 MBERT와 XLM-R가 한국어 사전 학습 모형보다 낮은 결과를 확인할 수 있었다. 이는 다국어 사전 학습 언어 모형이 기존의 학습된 구조를 이용하여 새로운 어휘로 매핑할 수는 있지만, 한국어와 같이 다른 언어를 수용하기 위해 문법, 어순과 같은 구조적인 부분의 변환을 학습하지는 않기 때문이라는 [11]의 주장에 근거하여 설명될 수 있다.

ROUGE 점수에 기반한 평가와 더불어 각 사전 학습 언어 모형에 의해 추출된 문장들에 대해, 해당 문장의 기존 문서에서의 문장 위치(첫 번째 문장, 두 번째 문장, …)에 대해 분석을 진행하였다.

그림 3은 신문 기사 데이터에 대한 기존 문서에서의 문장 위치에 따른 언어 모형별 추출 비율을 나타낸 그래프이다. 파란 막대는 실제 정답 요약 문장들의 위치를 나타내며, 전반적으로 문서 내에 고르게 분포하고 있다. 하지만 사전 학습된 모형들은 문서의 서두 부분에 집중해서 추출하는 경향성을

보이는 것으로 보아 Lead bias가 발생했음을 알 수 있다. 중요한 정보를 글의 서두에 배치하는 언론의 관례로 인해, 기사의 선행 문장들은 종종 핵심 정보를 포함하고 이는 Lead bias를 초래한다고 한다 [24]-[26].

이러한 문제는 사전 학습이 완료된 언어 모형 뿐만 아니라 사전 학습이 진행되지 않은 언어 모형에서도 발생하였다[10]. 따라서 Lead bias는 단순히 언어 모형의 문제가 아닌 기사 데이터가 가지는 특성 때문이라고 볼 수 있다. Lead bias를 해결하기 위해 문장의 순서를 무작위로 섞는 등의 문장 위치에 대한 편향을 줄이려는 연구들이 현재 진행 중에 있다 [25].

그림 4는 사설/잡지 데이터의 문장 위치에 따른 추출 비율을 나타낸다. 첫 번째 문장을 제외하고는 실제 정답 요약 문장의 위치는 문서 내에 고르게 퍼져 있다. 전반적으로 한국어 사전 학습 언어 모형과 다국어 사전 학습 언어 모형인 XLM-R이 실제 정답 요약 문장의 분포를 잘 따르는 것을 확인할 수 있다.

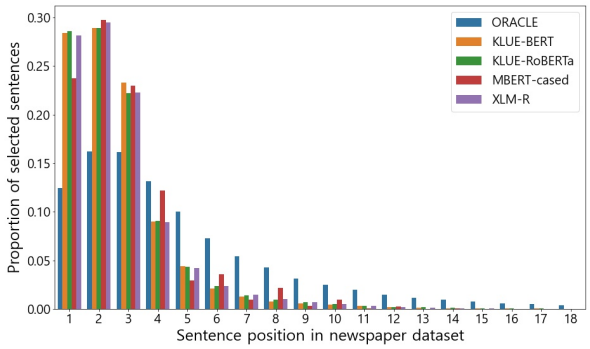

그림 3. 신문 기사 데이터의 문장 위치에 따른 추출 비율
Fig. 3. Extraction ratio according to the sentence position of newspaper data

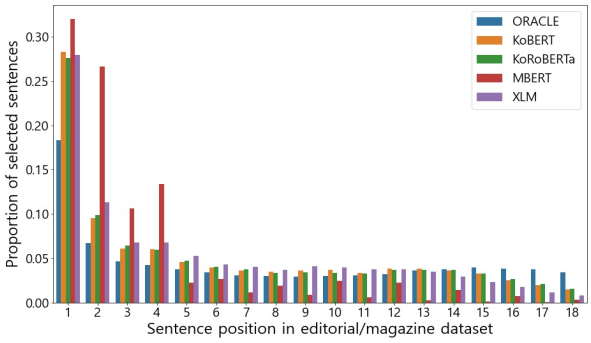

그림 4. 사설/잡지 데이터의 문장 위치에 따른 추출 비율
Fig. 4. Extraction ratio according to the sentence position of editorial/magazine data

표 6. 신문 기사 테스트 데이터의 추출 요약 결과 예시
Table 6. Example for extracted summary results of newspaper test data

Gold summary	100만달러 수출 계약 "패션 본고장서 인정" 현대백화점그룹 계열 패션전문기업 '한섬'이 북미·유럽 일상복(캐주얼)시장에 진출한다. 주력 일상복 브랜드인 '시스템'과 '시스템옴므'를 패션 본고장에 100만달러 이상 수출하기로 계약을 맺었다. 세계적인 패션브랜드 도약은 물론 '패션한류'에도 청신호가 켜진 셈이다.
Summary of KLUE-RoBERTa	100만달러 수출 계약 "패션 본고장서 인정" 현대백화점그룹 계열 패션전문기업 '한섬'이 북미·유럽 일상복(캐주얼)시장에 진출한다. 주력 일상복 브랜드인 '시스템'과 '시스템옴므'를 패션 본고장에 100만달러 이상 수출하기로 계약을 맺었다. 세계적인 패션브랜드 도약은 물론 '패션한류'에도 청신호가 켜진 셈이다.
Summary of KLUE-BERT	100만달러 수출 계약 "패션 본고장서 인정" 현대백화점그룹 계열 패션전문기업 '한섬'이 북미·유럽 일상복(캐주얼)시장에 진출한다. 주력 일상복 브랜드인 '시스템'과 '시스템옴므'를 패션 본고장에 100만달러 이상 수출하기로 계약을 맺었다. 한섬은 지난 1월 프랑스 톰그레이하운드 파리 매장에서 열린 1차 '시스템·시스템옴므' 2019 FW(가을·겨울)패션 단독 쇼룸(Showroom 공개전시회) 행사를 통해 11개국 20개 패션·유통업체와 홀세일(wholesale 도매) 계약을 맺었다고 24일 밝혔다.
Summary of XLM-R	주력 일상복 브랜드인 '시스템'과 '시스템옴므'를 패션 본고장에 100만달러 이상 수출하기로 계약을 맺었다. 세계적인 패션브랜드 도약은 물론 '패션한류'에도 청신호가 켜진 셈이다. 한섬은 지난 1월 프랑스 톰그레이하운드 파리 매장에서 열린 1차 '시스템·시스템옴므' 2019 FW(가을·겨울)패션 단독 쇼룸(Showroom 공개전시회) 행사를 통해 11개국 20개 패션·유통업체와 홀세일(wholesale 도매) 계약을 맺었다고 24일 밝혔다.
Summary of MBERT-cased	한섬은 1차 행사후 이달 4일부터 7일까지 같은 장소에서 2차 쇼룸 행사를 열고 또 다른 20여 개 업체와 계약 상담을 벌이고 있다. 앞서 한섬과 납품계약을 체결한 업체는 미국 '블루밍데일즈' 백화점, 캐나다 '라메종 사이먼스' 백화점, 이탈리아 하이엔드 패션편집숍 '안토니올리(Antonioli)', 홍콩 최대 패션편집숍 'I.T' 등 20곳이다. 이종호 한섬 브랜드지원담당 상무는 "세계 패션시장에 처음 뛰어든 신생 브랜드가 까다롭기로 유명한 유럽 편집숍 등 해외 패션·유통업체와 수출 계약까지 체결한 것은 이례적인 일"이라고 평가했다.

반면 MBERT는 실제 정답 요약 문장보다 더 많이 또는 덜 정보를 추출하는 모습을 보인다. 이는 ROUGE 점수와의 관계성으로 보아 다국어 사전 학습 언어 모형인 MBERT의 Undertrain 문제로 볼 수 있으며 문장 위치에 따른 추출 비율을 통해서도 해당 문제를 확인할 수 있었다.

4.4 통계적 검증

한국어 사전 학습 모형의 성능이 다국어 사전 학습 모형의 성능보다 통계적으로 유의미한 차이가 있는지 확인하였다. 한국어와 다국어 사전 학습 언어 모형에서 가장 성능이 좋았던 KLUE-RoBERTa와 XLM-R을 선별하여 검증을 하였다. 무작위로 KLUE-RoBERTa의 요약문 결과와 XLM-R의 요약문 결과를 추출한 후 ROUGE-2 점수를 지표로 선정하여 1000회 반복 수행하였다. 두 모형 결과 간에 통계적 유의미성을 검증하고자 이표본 t-검정을 진행하였다. 또한, 이표본 t-검정을 수행하기 전, Levene 의 등분산 검정을 통해 비교 집단과 대상 집단이 통계적으로 동일한 집단인지를 판별하였다.

표 8은 신문 기사에 대한 통계적 검증 결과이다. Levene의 등분산 검정 결과 p-value가 0.05보다 크므로 한국어 사전 학습 모형의 결과와 다국어 사전 학습 모형의 결과가 등분산을 따른다고 가정할 수 있다. 등분산을 따르는 두 모형의 결과에 대한 이표본 t-검정 결과는 p-value가 0.05보다 작고 95% 신뢰구간에서 0을 포함하지 않으므로 결과에 유의미한 차이가 있음을 확인하였다.

표 9는 사설/잡지에 대한 통계적 검증이다. 표 8의 결과와 마찬가지로 사설/잡지에 대한 두 모형에 대한 결과가 등분산을 따르며, 이표본 t-검정 결과 p-value가 0.05보다 작고 95% 신뢰구간에서 0을 포함하지 않으므로 유의미한 차이가 있다는 것을 확인하였다.

따라서, 한국어 문서에 대해서는 한국어 사전 학습 언어 모형이 다국어 사전 학습 언어 모형보다 좋다는 통계적인 검증 결과까지 확보하였다.

표 7. 사설/잡지 테스트 데이터의 추출 요약 결과 예시
Table 7. Example for extracted summary results of editorial/magazine test data

Gold summary	중국이 지난해 12월31일 세계보건기구에 '정체불명의 폐렴'으로 코로나19 발생을 공식 보고한 이후 100일을 갓 넘긴 10일 현재 전 세계 누적 확진자는 160만명을 넘어섰다. 기초과학연구원 RNA연구단 김빛내리 단장(서울대 생명과학부 교수)이 이끄는 공동 연구팀이 세계 최초로 '사스 코로나바이러스 2'의 유전자 지도를 완성해 공개한 것이다. 김 교수는 이번 유전자 지도가 코로나바이러스19의 증식원리를 상세히 알려줌으로써 더 정확한 진단키트와 새로운 치료 전략을 개발하는 데 기여할 것이라고 했다.
Summary of KLUE-RoBERTa	코로나19는 코로나바이러스의 7번째 인체 감염 사례이자 사스(SARS·중증급성호흡기증후군), 메르스(MERS·중동호흡기증후군)에 이은 3번째 중증 질환이다. 기초과학연구원 RNA연구단 김빛내리 단장(서울대 생명과학부 교수)이 이끄는 공동 연구팀이 세계 최초로 '사스 코로나바이러스 2'의 유전자 지도를 완성해 공개한 것이다. 김 교수는 이번 유전자 지도가 코로나바이러스19의 증식원리를 상세히 알려줌으로써 더 정확한 진단키트와 새로운 치료 전략을 개발하는 데 기여할 것이라고 했다.
Summary of KLUE-BERT	문제는 아직도 코로나19의 정점과 끝을 알 수 없는 '팬데믹'(세계적 대유행)이 이어지고 있다는 점이다. 기초과학연구원 RNA연구단 김빛내리 단장(서울대 생명과학부 교수)이 이끄는 공동 연구팀이 세계 최초로 '사스 코로나바이러스 2'의 유전자 지도를 완성해 공개한 것이다. 김 교수는 이번 유전자 지도가 코로나바이러스19의 증식원리를 상세히 알려줌으로써 더 정확한 진단키트와 새로운 치료 전략을 개발하는 데 기여할 것이라고 했다.
Summary of XLM-R	코로나19는 코로나바이러스의 7번째 인체 감염 사례이자 사스(SARS·중증급성호흡기증후군), 메르스(MERS·중동호흡기증후군)에 이은 3번째 중증 질환이다. 문제는 아직도 코로나19의 정점과 끝을 알 수 없는 '팬데믹'(세계적 대유행)이 이어지고 있다는 점이다. 기초과학연구원 RNA연구단 김빛내리 단장(서울대 생명과학부 교수)이 이끄는 공동 연구팀이 세계 최초로 '사스 코로나바이러스 2'의 유전자 지도를 완성해 공개한 것이다.
Summary of MBERT-cased	코로나바이러스는 포유류와 조류에 호흡기 질환을 일으키는 RNA 바이러스다. 1930년대 닭에서 처음 발견된 뒤 개·돼지·박쥐·돌고래 등으로 이어졌다. 사람 발병은 1960년대에 나타났다.

표 8. 신문 기사 테스트 데이터에 대한 통계적 검증
Table 8. Statistical validation of newspaper test data

Levene test		
Hypothesis	p- value	
$H_0: \mu_{KLUE-RoBERTa} = \mu_{XLM-R}$	p-value > 0.05	
$H_1: \mu_{KLUE-RoBERTa} \neq \mu_{XLM-R}$		
Two Sample t-test		
Hypothesis	p-value	95% C.I.
$H_0: \mu_{KLUE-RoBERTa} = \mu_{XLM-R}$	p-value < 0.05	[0.01206, 0.01210]
$H_1: \mu_{KLUE-RoBERTa} \neq \mu_{XLM-R}$		

표 9. 사설/잡지 테스트 데이터에 대한 통계적 검증
Table 9. Statistical validation of editorial/magazine test data

Levene test		
Hypothesis	p- value	
$H_0: \mu_{KLUE-RoBERTa} = \mu_{XLM-R}$	p-value > 0.05	
$H_1: \mu_{KLUE-RoBERTa} \neq \mu_{XLM-R}$		
Two Sample t-test		
Hypothesis	p-value	95% C.I.
$H_0: \mu_{KLUE-RoBERTa} = \mu_{XLM-R}$	p-value < 0.05	[0.00784393, 0.00786984]
$H_1: \mu_{KLUE-RoBERTa} \neq \mu_{XLM-R}$		

V. 결 론

본 연구는 한국어 문서 요약 연구를 위해 검증된 인공지능 통합 플랫폼인 AI hub에서의 데이터 세트를 활용하였다. 글의 특성이 다른 신문 기사와 사설 및 잡지 문서를 사용하였으며, 다양한 모델 간 비교 실험을 통해 문서 추출 요약에 대한 베이스라인을 제시한다. 문서의 편향성을 제거하기 위해 차별 및 혐오 표현이 담긴 일부 기사를 제외하고 기사와 무관한 내용을 텍스트 전처리를 통해 제거하였다. 사전 학습된 언어 모형의 성능에 대한 베이스라인의 적합성을 충족하고자, 동일한 구조의 사전 학습 언어 모형이지만 한국어로 학습된 모형과 다국어로 학습된 모형을 각각 2개씩 선정하여 비교 실험하였다. 한국어 문서에 대해서 한국어 사전 학습 언어 모형이 다국어 사전 학습 언어 모형을 최대 15.4% 가량 능가하는 모습을 보였다. 이를 바탕으로 단어 사전 구축 크기가 상당히 큰 다국어 사전 언어 모형이 한국어 개별 어휘에 대한 학습은 가능하나 어순과 같은 구조적인 측면을 학습하는 것에는 한계가 있다는 사실을 확인하였다. 더하여 MBERT 모형이 상당히 Undertrain되었다는 사실을 문서 요약 평가 지표인 ROUGE와 요약 문장으로 추출된 문장들의 위치에 대한 비율을 통해서도 확인할 수 있었다. 본 연구 결과를 통해 지금까지 미진했던 한국어 문서 요약에 대한 기본적인 베이스라인을 제시하는 동시에 앞으로의 한국어 NLP 과업을 진행할 시 효율적인 한국어 사전 학습 언어 모형 선정에도 도움이 될 수 있기를 기대한다.

References

[1] D. Shen, "Text Summarization BT - Encyclopedia of Database Systems", L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, pp. 3079-3083, 2009.

[2] J. M. Yoon, Y. J. Chung, and J. H. Lee, "Automatic Extractive Summarization of Newspaper Articles using Activation Degree of 5W1H", J. KIISE, SA, Vol. 31, No. 4, pp. 505-515, Apr. 2004.

[3] Y. Liu, "Fine-tune BERT for extractive summarization", arXiv Prepr. arXiv1903.10318,

Mar. 2019. https://doi.org/10.48550/arXiv.1903.10318.

[4] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher, "Evaluating the Factual Consistency of Abstractive Text Summarization", In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, pp. 9332-9346, Association for Computational Linguistics, Nov. 2020. http://dx.doi.org/10.18653/v1/2020.emnlp-main.750.

[5] Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig, "GSum: A General Framework for Guided Neural Abstractive Summarization", In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, pp. 4830-4842, Association for Computational Linguistics, Jun. 2021. http://dx.doi.org/10.18653/v1/2021.naacl-main.384.

[6] Jaewon Jeaon, Hyunsun Hwang, and Changki Lee, "Two-step Document Summarization using Deep Learning and Maximal Marginal Relevance", Korea Information Processing Society, pp. 347-349, Oct. 2019.

[7] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents", 2017.

[8] Gyoungho Lee, Yohan Park, and Kongjoo Lee, "Building a Korean Text Summarization Dataset Using News Articles of Social Media", Korea Information Processing Society. Transactions on Software and Data Engineering, Vol. 9, No. 8, pp. 251-258, Aug. 2020. https://doi.org/10.3745/KTSDE.2020.9.8.251.

[9] AI Hub, "Manual for text summarization dataset", 2021. https://aihub.or.kr/aidata/8054. [accessed: Jan. 21, 2022]

[10] Yang Liu and Mirella Lapata, "Text Summarization with Pretrained Encoders", In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 3730-3740, Association for Computational Linguistics.19, Aug. 2019. https://doi.org/10.48550/arXiv.1908.08345.

[11] Telmo Pires, Eva Schlinger, and Dan Garrette, "How Multilingual is Multilingual BERT?", In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 4996-5001, Association for Computational Linguistics, Jul. 2019. http://dx.doi.org/10.18653/v1/P19-1493.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale", In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, pp. 8440-8451, Association for Computational Linguistics, Jul. 2020. http://dx.doi.org/10.18653/v1/2020.acl-main.747.

[13] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training", 2018.

[14] Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung, "BanditSum: Extractive Summarization as a Contextual Bandit", In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 3739-3748, Association for Computational Linguistics, Oct.-Nov. 2018. http://dx.doi.org/10.18653/v1/D18-1409.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In Proceedings of the 2019 Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186, Association for Computational Linguistics, Jun. 2019. http://dx.doi.org/10.18653/v1/N19-1423.

[16] E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep Contextualized Word Representations", In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers), pNew Orleans, Louisiana, pp. 2227-2237, Association for Computational Linguistics, Jun. 2018. http://dx.doi.org/10.18653/v1/N18-1202.

[17] Y. Liu, "Roberta: A robustly optimized bert pretraining approach", Conference paper, ICLR Jul. 2019. https://doi.org/10.48550/arXiv.1907.11692.

[18] Shashi Narayan, Shay B. Cohen, and Mirella Lapata, "Ranking Sentences for Extractive Summarization with Reinforcement Learning", In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers), New Orleans, Louisiana, pp. 1747-1759, Association for Computational Linguistics, Jun. 2018. http://dx.doi.org/10.18653/v1/N18-1158.

[19] Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao, "Neural Document Summarization by Jointly Learning to Score and Select Sentences", In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Melbourne, Australia, pp. 654-663, Association for Computational Linguistics, Jul. 2018. http://dx.doi.org/10.18653/v1/P18-1061.

[20] S. Park, "KLUE: Korean Language Understanding Evaluation", arXiv Prepr. arXiv2105. 09680, Nov. 2021. https://doi.org/10.48550/arXiv. 2105.09680.

[21] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", In Text Summarization Branches Out, Barcelona, Spain, pp. 74–81, Association for Computational Linguistics, Jul. 2004.

[22] Tsutomu Hirao, Masaaki Nishino, and Masaaki Nagata, "Oracle Summaries of Compressive Summarization", In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers), Vancouver, Canada, pp. 275-280, Association for Computational Linguistics, Jul. 2017. http://dx.doi.org/10.18653/v1/P17-2043.

[23] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization", BT - 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 2019. https://doi.org/10.48550/arXiv.1711.05101.

[24] Chris Kedzie, Kathleen McKeown, and Hal Daumé III, "Content Selection in Deep Learning Models of Summarization", In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, pp. 1818-1828, Association for Computational Linguistics, Oct.-Nov. 2018. http://dx.doi.org/10.18653/v1/D18-1208.

[25] Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis, "Countering the Effects of Lead Bias in News Summarization via Multi-Stage Training and Auxiliary Losses", In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 6019–6024, Association for Computational Linguistics, Nov. 2019. http://dx.doi.org/10.18653/v1/D19-1620.

[26] Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang, "Leveraging Lead Bias for Zero-shot Abstractive News Summarization", Proceedings of the 44th International ACM SIGIR Conference on Research

and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, pp. 1462-1471, Jul. 2021. https://doi.org/10.1145/3404835.3462846.

저자소개

박 재 언 (Jae-Eon Park)

2020년 8월 : 국민대학교 빅데이터경영통계학과(졸업)
2020년 9월 ~ 현재 : 고려대학교 산업경영공학과(석사과정)
관심분야 : 인공지능, 자연어처리, 문서 요약

김 지 호 (Ji-Ho Kim)

2015년 8월 : 서울과학기술대학교 글로벌융합산업공학과(졸업)형
2015년 9월 ~ 현재 : 고려대학교 산업경영공학부(석박사통합과정)
관심분야 : 인공지능, 자연어처리, 비즈니스 인텔리전스

이 홍 철 (Hong-Chul Lee)

1983년 : 고려대학교 산업공학부(학사)
1986년 ~ 1988년 : University of Texas Arlington, Industrial Engineering (M.S.)
1988년 ~ 1993년 : Texas A&M University, Industrial Engineering (Ph.D.)
1995년 ~ 1996년 : Research Committee ,Production and technology Institute of Korea University
1993년 ~ 1994년 : The University of Iowa, Post. Doc
1986년 ~ 1993년 : Research and Teaching Assistant, Texas A&M University
1984년 ~ 1986년 : Mathematics lecturer Korea Military Academy
1996년 ~ 현재 : 고려대학교 산업경영공학부 교수
관심분야 : 인공지능, 생산공학시스템, 시뮬레이션